Rough draft for a White Paper, LAT Bauerdick et al

This white paper is to make the case for a new class of scientific tools that support complex collaborative scientific workflows and data management within distributed computing environments.

Science has become a vastly more complex endeavor. Scientific collaborations are becoming not only larger, but also more distributed and more diverse. The scientific community is responding to the challenge by creating global collaborations, petascale data infrastructures and internationally distributed computing environments.

As an example, the science community at the Large Hadron Collider (LHC) is utilizing a community-wide application grid linking tens of thousands of computers and petabytes of storage to provide globally distributed computational systems and access to the complex data collections. LHC science depends on the global grid infrastructure and collaborative analysis environments.

However the usability of these vast computing resources is still poor affecting the productivity of researchers in analyzing the large data samples for discoveries. There are only rudimentary tools to share information, data and metadata, and workflows within working groups that are typically distributed across several institutions and often across different counties. Existing tools are primitive and often do not provide secure sharing or fine-grained access control, there is little or no integration between the data-access and data reduction workflows and the production of final diagrams and results and tracking of data provenance is often not supported or difficult, and very little support exists for pulling together information assembled from a large number of disjoint individual work flows and individual contributions.

This is to be contrasted with the level of tools that a “knowledge worker” or engineer can rely on in a local workspace environment, like on single desktop or small workgroup cluster. On the desktop there exist well developed tools and approaches that support complex workflows, large amount of data and metadata, safe sharing of information, and support for collaborative work. An example for a highly functional tool in the local environment would be a video editing application that provides very sophisticated and integrated tools for workflows with complexity not so different from those required in scientific data analysis. These desktop application environments provide well-developed intuitive paradigms to support these workflows, to manage the associated data, and to hide underlying complexity. These tools often rely on concepts that have a sound CS foundation, like databases, while much that exist in the distributed environment is ad-hoc.

The next step should be to research, prototype and deploy the user level tools for the distributed computing environment that will enable scientific collaborators to

work together as co-located peers, to share resources, and to provide a "workspace" environment in which researchers can encapsulate all the information regarding their scientific workflows and

Such a "workspace" paradigm would provide ways to share and configure, to allocate resources across the distributed environment, to encapsulate all relevant workflow and data management information, to supports data provenance information, to keeps information about state, to bind together contributions from different and distributed work groups, to make sharing of information and data easy lowering the threshold for deep collaboration, to makes information discoverable across large collaborations, to expose scientific processes to the public.

The following picture (developed by Ruth Pordes) illustrates elements of such a set of tools supporting scientific workspaces.